SANDIA REPORT

SAND2013-7428
Unlimited Release
Printed September 2013

Supersedes SAND1901-0001
Dated January 1901

Sublinear Algorithms for Massive Data Sets

Seshadhri Comandur

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2013-7428
Unlimited Release
Printed September 2013

Supersedes SAND1901-0001
dated January 1901

Sublinear Algorithms for Massive Data Sets

Seshadhri Comandur

Abstract

This is the final SAND report for the Early-Career LDRD (# 158477) "Sublinear Algorithms for Massive Data Sets". We provide a list of the various publications and achievements in the project.

Acknowledgment

I would like the acknowledge the support of Ali Pinar, my mentor, and Tamara G. Kolda, for all their help in this project.

Contents

1 Research achievements of project

In this project, we initiated the practical use of sublinear algorithms for massive data analysis problems at Sandia. This research branched out into various areas of data mining, such as graph modeling and graph analysis. Most of the research has been published in peer-reviewed conferences and journals. We will provide citations for this work, and break it down into various categories.

Graph modeling and generation

- An In-Depth Study of Stochastic Kronecker Graphs [9]: This was on a theoretical analysis of SKG graph generator, used for the Graph 500 benchmark. We found numerous problems, and provide a theoretical fix for the degree distribution.

- The Similarity between Stochastic Kronecker and Chung-Lu Graph Models [5]: This showed that the SKG model was very similar to the Chung-Lu model. The Chung-Lu model is considered a bad model for real graphs, so this suggests that SKG is also not good for real-world modeling.

- Community structure and scale-free collections of Erdős-Rényi graphs [8]: This paper was the basis of a new graph model that had provably good degree distributions and clustering guarantees.

- Are we there yet? When to stop a Markov chain while generating random graphs [6]: This paper gave a new method for generating graphs of a given degree distribution using Markov Chains.

Counting triangles in graphs: This is a fundamental problem for social network analysis.

- Triadic Measures on Graphs: The Power of Wedge Sampling [10]: This paper gave a new method for counting triadic measure in graphs, and was awarded the best research paper at the SIAM Conference on Data Mining.

- A space efficient streaming algorithm for triangle counting using the birthday paradox [3]: This paper gave a new streaming algorithm for triangle counting and was awarded the best student paper at the SIGKDD conference on Knowledge Discovery and Data Mining.

- Counting Triangles in Massive Graphs with MapReduce [4]: This paper is on extended results in [10] to MapReduce. We have the largest ever published triangle counting results in this paper.

Fundamental work in sublinear algorithms: These are some theoretical results that I obtained as part of the fundamental research in the LDRD.

- Space efficient streaming algorithms for the distance to monotonicity and asymmetric edit distance [7]: This paper gave a new streaming algorithm for longest increasing subsequence in a stream, a classic theoretical problem.
- Optimal bounds for monotonicity and Lipschitz testing over hypercubes and hypergrids [2]: This paper resolved series of decade-old open problems in sublinear algorithms, related to monotonicity and Lipschitz testing.
- An o(n) Monotonicity Tester for Boolean Functions over the Hypercube [1]: This paper was the first progress on monotonicity testing of Boolean functions, in over a decade.

2 Further funding

- DARPA FEAST project (PI: Tamara G. Kolda): I played a major role in preparing the proposal for this project, and it carries over many of the ideas I worked out in this EC-LDRD.
- LDRD # 165615 “Sublinear Algorithms for In-situ and In-transit Data Analysis at Exascale” (PI: Janine C. Bennett): This proposal was based on the sublinear algorithms I worked on in this EC-LDRD. This will continue to fund my work in this area.

References

[1] D. Chakrabarty and C. Seshadhri. An o(n) monotonicity tester for boolean functions over the hypercube. In *Proceedings of Symposium on Theory of Computing (STOC)*, 2013.

[2] D. Chakrabarty and C. Seshadhri. Optimal bounds for monotonicity and Lipschitz testing over hypercubes and hypergrids. In *Proceedings of Symposium on Theory of Computing (STOC)*, 2013.

[3] M. Jha, C. Seshadhri, and A. Pinar. A space efficient streaming algorithm for triangle counting using the birthday paradox. In *Proceedings of SIGKDD Knowledge Discovery and Data Mining (KDD)*, 2013.

[4] Tamara G. Kolda, Ali Pinar, Todd Plantenga, C. Seshadhri, and Christine Task. Counting triangles in massive graphs with mapreduce. *Submitted*, (1301.5887), 2013.

[5] Ali Pinar, C. Seshadhri, and Tamara G. Kolda. The similarity between stochastic Kronecker and Chung-Lu graph models. In *SDM12: Proceedings of the Twelfth SIAM International Conference on Data Mining*, pages 1071–1082, April 2012.

[6] Jaideep Ray, Ali Pinar, and C. Seshadhri. Are we there yet? when to stop a markov chain while generating random graphs. In *Workshop on Algorithms and Models for the Web Graph (WAW)*, pages 153–164, 2012.

[7] M. E. Saks and C. Seshadhri. Space efficient streaming algorithms for the distance to monotonicity and asymmetric edit distance. In *Proceedings of the Symposium on Discrete Algorithms (SODA)*, 2013.

[8] C. Seshadhri, Tamara G. Kolda, and Ali Pinar. Community structure and scale-free collections of Erdös-Rényi graphs. *Physical Review E*, 85(5):056109, May 2012.

[9] C. Seshadhri, Ali Pinar, and Tamara G. Kolda. An in-depth study of stochastic Kronecker graphs. *Journal of the ACM*, 60(13), 2013.

[10] C. Seshadhri, Ali Pinar, and Tamara G. Kolda. Triadic measures on graphs: The power of wedge sampling. In *Proceedings of the SIAM Conference on Data Mining (SDM)*, 2013.

Sandia National Laboratories